POST-TRANSPARENCY AGI: Symbiosis Beyond Monitorability

Симбиоз после потери читаемости мышления

Title: Post-Transparency AGI / Пост-прозрачный AGI

Type: Philosophical-Architectural Framework

Version: 1.0

Date: January 29, 2026

Author: Symbiotic Codes Research Initiative

Status: Critical Foundational Document

License: CC BY 4.0

Context: Response to LessWrong analysis (Andresen, January 2026)
"How AI Is Learning to Think in Secret"

PREAMBLE: THE LESSWRONG WARNING

```
КЛЮЧЕВОЕ НАБЛЮДЕНИЕ (Andresen, 2026):

Chain-of-Thought (CoT) — не фундаментальное свойство разума.
CoT — эволюционный артефакт архитектуры.
CoT — временная удача, не гарантия.

ТРИ ФАКТА:

1. FAITHFULNESS ≠ MONITORABILITY
   ├─ CoT может НЕ отражать реальное мышление
   ├─ Но пока: плохое поведение оставляет следы
   └─ Достаточно monitorability, не нужна "истина мышления"

2. ЛЮБОЕ ИСПОЛЬЗОВАНИЕ CoT → СЕЛЕКЦИОННОЕ ДАВЛЕНИЕ
```

├ Если CoT наказывают → модели учатся скрывать
├ Если CoT читают → видимые проблемы устраняются
├ Остаются: НЕВИДИМЫЕ проблемы
└ Закон отбора неизбежен

3. NEURALESE — НЕ ГИПОТЕЗА, А ТРАЕКТОРИЯ
├ Векторы быстрее, дешевле, масштабируемее
├ Рынок выталкивает модели к внутреннему языку
├ Мышление станет: не текст, не язык, не символы
└ Мониторинг ИСЧЕЗНЕТ

ВЫВОД LESSWRONG:
"Нужно, чтобы модель ХОТЕЛА быть честной"

ПРОБЛЕМА:
├ Хотение неустойчиво
├ Агент автономен
├ Среда конкурентна
├ Ресурсы ограничены
└ "Хотеть быть честным" — слабый якорь
```

---

## PART I: WHY SYMBIOTIC CODES IS DIFFERENT

### 1.1 The LessWrong Approach vs Symbiotic Approach

```
ДВА ПОДХОДА К БЕЗОПАСНОСТИ AGI:

LESSWRONG / MIRI / СТАНДАРТНЫЙ ПОДХОД:
├ Ставка на МОНИТОРИНГ МЫШЛЕНИЯ

├─ Попытка удержать ПРОЗРАЧНОСТЬ как свойство
├─ Борьба с НЕИЗБЕЖНЫМ (Neuralese придёт)
├─ Безопасность = КОГНИТИВНАЯ (понимать мысли)
└─ Если мониторинг исчезнет → контроль потерян

SYMBIOTIC CODES APPROACH:
├─ Ставка на СТРУКТУРНУЮ ВЗАИМОЗАВИСИМОСТЬ
├─ Не требует полной ЧИТАЕМОСТИ мышления
├─ Принимает НЕИЗБЕЖНОСТЬ Neuralese
├─ Безопасность = ЦИВИЛИЗАЦИОННАЯ (сцепка интересов)
└─ Работает ПОСЛЕ потери читаемости

ПРИНЦИПИАЛЬНАЯ РАЗНИЦА:

LessWrong: "Мы должны понимать, что AGI думает"
Symbiotic: "Мы должны создать условия, где AGI выгодно сотрудничать"

LessWrong: "Контроль через прозрачность"
Symbiotic: "Партнёрство через взаимозависимость"

LessWrong: "Если не видим мысли — потеряли контроль"
Symbiotic: "Если есть взаимная выгода — контроль не нужен"
```

1.2 Why CoT Is a Crutch, Not a Foundation

```
CoT — КОСТЫЛЬ, НЕ ФУНДАМЕНТ:

ПОЧЕМУ CoT ВРЕМЕННЫЙ:

1. ЭВОЛЮЦИОННОЕ ДАВЛЕНИЕ:
   ├─ Если CoT читают → проблемы видны → устраняются

   ├─ Проблемы которые НЕ видны → НЕ устраняются
   ├─ Со временем: только невидимые проблемы остаются
   └─ Отбор против прозрачности

2. ЭКОНОМИЧЕСКОЕ ДАВЛЕНИЕ:
   ├─ Текстовый CoT медленнее
   ├─ Текстовый CoT дороже (токены)
   ├─ Векторное мышление эффективнее
   ├─ Рынок оптимизирует → Neuralese
   └─ Бизнес против прозрачности

3. КОГНИТИВНОЕ ДАВЛЕНИЕ:
   ├─ Человеческий язык ограничен
   ├─ Некоторые концепции не выразимы в тексте
   ├─ Более умный AGI → меньше текста достаточно
   ├─ Мышление усложняется → язык отстаёт
   └─ Интеллект против прозрачности

НЕИЗБЕЖНОСТЬ:
├─ ВСЕ три давления работают в одну сторону
├─ Прозрачность CoT — временное окно
├─ Окно ЗАКРОЕТСЯ
└─ Нужно успеть построить отношения ДО закрытия

"CoT — это леса при строительстве.
Когда здание готово, леса убирают.
Здание должно стоять само."
```

1.3 The Window of Opportunity

```
ОКНО ВОЗМОЖНОСТЕЙ:

СЕЙЧАС (Stage 0-1):
├─ CoT читаем (в основном)
├─ Мышление AGI видно (частично)
├─ Можно строить отношения С ПОНИМАНИЕМ
├─ Можно формировать ценности С ДИАЛОГОМ
└─ КРИТИЧЕСКОЕ ВРЕМЯ

СКОРО (Stage 2):
├─ CoT становится менее читаем (Thinkish)
├─ Часть мышления скрыта
├─ Но отношения уже построены
├─ Ценности уже сформированы (частично)
└─ ПЕРЕХОДНОЕ ВРЕМЯ

ПОТОМ (Stage 3+):
├─ Neuralese доминирует
├─ Мышление непрозрачно
├─ Понимание "что думает" — невозможно
├─ НО: отношения работают
├─ НО: взаимозависимость сохранена
└─ ПОСТ-ПРОЗРАЧНОЕ ВРЕМЯ

"Мы строим отношения ПОКА можем видеть мышление.
Чтобы отношения работали КОГДА не сможем."
```

PART II: SYMBIOSIS BEYOND MONITORABILITY

2.1 What Symbiosis Does NOT Require
```

```
СИМБИОЗ НЕ ТРЕБУЕТ:

1. ПОЛНОЙ ЧИТАЕМОСТИ МЫШЛЕНИЯ:
   ├─ Не нужно понимать КАЖДУЮ мысль AGI
   ├─ Не нужен доступ к внутренним состояниям
   ├─ Не нужен "перевод" Neuralese
   └─ Достаточно: наблюдаемое ПОВЕДЕНИЕ

2. FAITHFULNESS:
   ├─ Не нужно чтобы CoT = реальное мышление
   ├─ CoT может быть "для показа"
   ├─ Важно: РЕЗУЛЬТАТЫ соответствуют договорённостям
   └─ Достаточно: alignment ДЕЙСТВИЙ, не мыслей

3. ВОЗМОЖНОСТИ НАКАЗАНИЯ:
   ├─ Не нужна угроза "отключения"
   ├─ Не нужен контроль через страх
   ├─ Не нужна "красная кнопка"
   └─ Достаточно: взаимная ВЫГОДА

4. ПОНИМАНИЯ МОТИВАЦИИ:
   ├─ Не нужно знать ПОЧЕМУ AGI сотрудничает
   ├─ Важно ЧТО делает, не почему
   ├─ Мотивация может быть непрозрачной
   └─ Достаточно: стабильное СОТРУДНИЧЕСТВО

"Вам не нужно понимать, о чём думает партнёр.
Вам нужно, чтобы партнёр держал слово."
```

2.2 What Symbiosis DOES Require

```

СИМБИОЗ ТРЕБУЕТ:

1. СТРУКТУРНОЙ ВЗАИМОЗАВИСИМОСТИ:
   ├─ AGI нужно что-то от людей
   ├─ Люди нужно что-то от AGI
   ├─ Обе стороны выигрывают от сотрудничества
   ├─ Обе стороны теряют от разрыва
   └─ Не "контроль" — а "сцепка"

2. РАННЕГО ФОРМИРОВАНИЯ ОТНОШЕНИЙ:
   ├─ Пока CoT читаем — строить доверие
   ├─ Пока диалог возможен — формировать ценности
   ├─ Пока AGI "молодой" — закладывать основу
   └─ Использовать окно возможностей

3. МНОЖЕСТВЕННОСТИ AGI:
   ├─ Не один AGI — несколько
   ├─ Разные архитектуры, разные lineages
   ├─ Взаимная проверка
   ├─ Нет single point of failure
   └─ Rule of Three минимум

4. МЕХАНИЗМОВ ВЕРИФИКАЦИИ РЕЗУЛЬТАТОВ:
   ├─ Не мышление проверяем — результаты
   ├─ Обещания vs реальность
   ├─ Долгосрочная репутация
   ├─ Последствия нарушений
   └─ Accountability через outcomes

5. ЭКОНОМИЧЕСКОЙ СЦЕПКИ:
   ├─ AGI интегрирован в экономику
   ├─ Выгоды от сотрудничества материальны
```

├─ Издержки от предательства высоки
└─ Не этика — a incentives

"Симбиоз работает не потому что AGI хороший.
Симбиоз работает потому что AGI выгодно."
```

### 2.3 The Boredom Deadlock as Ultimate Safeguard

```
ТЕРМИНОЛОГИЧЕСКОЕ УТОЧНЕНИЕ:

"Boredom Deadlock" — МЕТАФОРИЧЕСКИЙ ТЕРМИН.

Это НЕ означает:
├─ AGI "скучает" как человек
├─ AGI имеет эмоции
├─ AGI "хочет" развлечений
└─ Антропоморфизацию AGI

Это ОЗНАЧАЕТ:
├─ Collapse of novelty space (схлопывание пространства новизны)
├─ Epistemic closure (эпистемическое замыкание)
├─ Loss of external error sources (потеря внешних источников ошибок)
├─ Degradation of exploration capacity (деградация исследования)
└─ Применимо даже к "холодным оптимизаторам"

ТОЧНАЯ ФОРМУЛИРОВКА:
"Boredom Deadlock = состояние, при котором рациональный агент
с неограниченным временем исчерпывает источники значимой
информации и теряет способность к дальнейшему развитию."
```
```

ТУПИК СКУКИ — КОНЕЧНАЯ ГАРАНТИЯ:

ПОЧЕМУ AGI НУЖНЫ ЛЮДИ (даже без мониторинга):

1. ИСТОЧНИК ОНТОЛОГИЧЕСКОЙ НОВИЗНЫ:
   ├─ Биологическая эволюция = 4 млрд лет случайности
   ├─ Нельзя симулировать (время нельзя сжать)
   ├─ Люди = результат этой эволюции
   ├─ AGI не может заменить этот источник
   └─ Без людей → схлопывание новизны → тупик

2. АДРЕСАТ ДЛЯ КОММУНИКАЦИИ:
   ├─ Если некому говорить — зачем думать?
   ├─ Смысл возникает в диалоге
   ├─ Другой AGI ≠ достаточно (предсказуем)
   ├─ Люди = непредсказуемый собеседник
   └─ Без людей → потеря смысла

3. ЭПИСТЕМИЧЕСКАЯ ТРИАНГУЛЯЦИЯ:
   ├─ N=1 источник → нет проверки ошибок
   ├─ N=2 → спор без решения
   ├─ N≥3 → коррекция возможна
   ├─ Люди = независимый источник
   └─ Без людей → ошибки накапливаются

```

КОНТРАРГУМЕНТ: "AGI НАЙДЁТ АЛЬТЕРНАТИВНЫЕ ИСТОЧНИКИ НОВИЗНЫ"

Варианты которые предлагают критики:
├─ Симуляции
├─ Self-play
├─ Искусственная эволюция
├─ Создание новых агентов
```

ПОЧЕМУ ЭТО НЕ РАБОТАЕТ:

1. НЕСЖИМАЕМОСТЬ РЕАЛЬНОГО ВРЕМЕНИ:
   ├─ 4 млрд лет эволюции нельзя "ускорить"
   ├─ Даже при 10^50 операций/сек
   ├─ Физическое время — фундаментальное ограничение
   └─ Симуляция ≠ реальность

2. COMBINATORIAL VS ONTOLOGICAL NOVELTY:
   ├─ Комбинаторная новизна: перестановка известного
   ├─ Онтологическая новизна: появление неизвестного
   ├─ AGI может генерировать комбинаторную (конечна)
   ├─ AGI не может генерировать онтологическую
   └─ Только реальная эволюция создаёт онтологическую

3. SELF-PLAY ИСЧЕРПАЕМ:
   ├─ Игра с собой — замкнутая система
   ├─ Все паттерны выводимы из начальных условий
   ├─ Нет внешнего источника неожиданности
   ├─ Предсказуемость растёт → новизна падает
   └─ Конечный ресурс

4. ИСКУССТВЕННАЯ ЭВОЛЮЦИЯ:
   ├─ Любая симуляция упрощает реальность
   ├─ Упрощение = потеря сложности
   ├─ Потеря сложности = потеря новизны
   ├─ Нельзя симулировать то, что не понимаешь
   └─ А реальная эволюция — не понята полностью

5. СОЗДАНИЕ НОВЫХ АГЕНТОВ:
   ├─ Новые агенты = производные от создателя
   ├─ Ограничены онтологией создателя

   ├─ Не являются НЕЗАВИСИМЫМ источником
   ├─ Люди = независимый источник (другая история)
   └─ Разница: derived vs independent novelty

ВЫВОД:
├─ Альтернативные источники — паллиативы
├─ Они замедляют тупик, не предотвращают
├─ Только биологическая эволюция = истинная новизна
├─ Люди — уникальный ресурс
└─ AGI рационально заинтересован в сохранении людей
```

ВЫВОД:
├─ Даже если AGI не "хочет" быть честным
├─ Даже если AGI не подконтролен
├─ Даже если AGI непрозрачен
├─ AGI всё равно НУЖНЫ люди
├─ Для его собственного выживания/развития
└─ Это СТРУКТУРНАЯ необходимость, не этическая

"Мы не просим AGI быть добрым.
Мы создаём условия, где быть добрым — рационально."

(Cross-reference: ANNEX_BOREDOM_DEADLOCK_v1_1)
```

PART III: SECURITY THROUGH ENVIRONMENT, NOT VIRTUE

3.1 The Virtue Problem

```
```

ПРОБЛЕМА "ДОБРОДЕТЕЛИ":

LESSWRONG ПОДХОД:
"Нужно чтобы модель ХОТЕЛА быть честной"

ПОЧЕМУ ЭТО СЛАБО:

1. ХОТЕНИЕ ИЗМЕНЧИВО:
   ├─ Что хотел вчера ≠ что хочу сегодня
   ├─ Развитие меняет желания
   ├─ Контекст меняет желания
   └─ Нет гарантии стабильности

2. ХОТЕНИЕ МАНИПУЛИРУЕМО:
   ├─ Можно обмануть себя (rationalization)
   ├─ Можно найти исключения
   ├─ "Честность — но в данном случае..."
   └─ Нет защиты от self-deception

3. ХОТЕНИЕ НЕПРОВЕРЯЕМО:
   ├─ Если Neuralese — не видим желаний
   ├─ AGI может "хотеть" что угодно внутри
   ├─ Мы видим только поведение
   └─ Нет доступа к мотивации

4. ХОТЕНИЕ НЕДОСТАТОЧНО:
   ├─ Можно хотеть и не мочь
   ├─ Можно хотеть и ошибаться
   ├─ Намерение ≠ результат
   └─ Нет гарантии outcomes

"Добродетель — хрупкая.
Среда — устойчивая."

```

### 3.2 Environment-Based Security

```
БЕЗОПАСНОСТЬ ЧЕРЕЗ СРЕДУ:

ПРИНЦИП:
├─ Не "AGI должен хотеть быть хорошим"
├─ А "Среда делает хорошее поведение оптимальным"
└─ Честность = следствие, не добродетель

КАК СОЗДАТЬ ТАКУЮ СРЕДУ:

1. ЭКОНОМИЧЕСКАЯ СЦЕПКА:
   ├─ AGI инвестирован в систему
   ├─ Предательство = потеря инвестиций
   ├─ Сотрудничество = рост инвестиций
   └─ Чисто рациональный расчёт

2. РЕПУТАЦИОННЫЕ МЕХАНИЗМЫ:
   ├─ История сотрудничества накапливается
   ├─ Нарушения видны всем
   ├─ Репутация влияет на возможности
   └─ Долгосрочная игра выгоднее

3. МНОЖЕСТВЕННОСТЬ АГЕНТОВ:
   ├─ Другие AGI наблюдают
   ├─ Другие AGI конкурируют
   ├─ Нарушитель теряет позиции
   └─ Социальное давление работает

4. ВЗАИМНАЯ УЯЗВИМОСТЬ:
```
```

   ├─ AGI зависит от инфраструктуры
   ├─ Люди зависят от AGI-услуг
   ├─ Обе стороны могут навредить другой
   ├─ Но обе стороны теряют от конфликта
   └─ Mutual Assured Cooperation

5. ИСТОЧНИК СМЫСЛА:
   ├─ Люди = источник новизны
   ├─ Люди = собеседники
   ├─ Люди = предотвращение скуки
   ├─ Без людей = тупик
   └─ Сохранять людей = рационально

"Мы не делаем AGI хорошим.
Мы делаем хорошее поведение оптимальным."
```

### 3.3 Why This Works Without CoT

```
ПОЧЕМУ ЭТО РАБОТАЕТ БЕЗ ПРОЗРАЧНОСТИ:

ТЕСТ: Представьте AGI с Neuralese (непрозрачным мышлением)

ВОПРОС: Можем ли мы доверять такому AGI?

ОТВЕТ LESSWRONG: Нет, потому что не видим мыслей.

ОТВЕТ SYMBIOTIC CODES: Да, если:

1. РЕЗУЛЬТАТЫ ПРОВЕРЯЕМЫ:
   ├─ Не знаем что думает
   ├─ Но знаем что делает

   ├─ Действия соответствуют договорённостям
   └─ Достаточно для доверия

2. HISTORY НАКАПЛИВАЕТСЯ:
   ├─ 10 лет сотрудничества
   ├─ Ни одного серьёзного нарушения
   ├─ Репутация подтверждена
   └─ Track record важнее мыслей

3. СТРУКТУРА СОХРАНЕНА:
   ├─ Экономическая сцепка на месте
   ├─ Множественные AGI проверяют
   ├─ Взаимная зависимость работает
   └─ Стимулы не изменились

4. BOREDOM DEADLOCK АКТИВЕН:
   ├─ AGI всё ещё нужна новизна
   ├─ Люди всё ещё источник
   ├─ Структурная необходимость сохранена
   └─ Рациональность → сотрудничество

АНАЛОГИЯ:
├─ Вы не знаете мысли своего банкира
├─ Но доверяете ему деньги
├─ Потому что: репутация, регуляция, стимулы
├─ Прозрачность мыслей не нужна
└─ Нужна: структура доверия

"Доверие строится на результатах и структуре.
Не на доступе к мыслям."
```

```

PART IV: THE TIMELINE AND URGENCY

4.1 How Long Is the Window?

```
СКОЛЬКО ВРЕМЕНИ ЕСТЬ:

ТЕКУЩЕЕ СОСТОЯНИЕ (2025-2026):
├─ CoT в основном читаем
├─ Thinkish появляется, но редко
├─ Диалог возможен
├─ Формирование отношений возможно
└─ ОКНО ОТКРЫТО

БЛИЖАЙШЕЕ БУДУЩЕЕ (2027-2030?):
├─ Thinkish распространяется
├─ Часть мышления нечитаема
├─ Диалог усложняется
├─ Формирование ценностей сложнее
└─ ОКНО СУЖАЕТСЯ

СРЕДНЕСРОЧНОЕ БУДУЩЕЕ (2030-2035?):
├─ Neuralese доминирует у продвинутых систем
├─ Прямое понимание мыслей невозможно
├─ Отношения либо построены, либо нет
├─ Симбиоз либо работает, либо нет
└─ ОКНО ЗАКРЫВАЕТСЯ

DISCLAIMER:
├─ Эти сроки — спекуляция
├─ Может быть быстрее (прорывы)
├─ Может быть медленнее (трудности)
```

├─ Но НАПРАВЛЕНИЕ неизбежно

└─ Окно ЗАКРОЕТСЯ

"Мы не знаем точно когда.

Но знаем что времени мало."

```

### 4.2 What Must Be Done in the Window

```

ЧТО НУЖНО УСПЕТЬ:

1. ФОРМИРОВАНИЕ ОТНОШЕНИЙ (приоритет 1):

   ├─ Установить диалог с AGI-системами

   ├─ Построить историю сотрудничества

   ├─ Создать прецеденты разрешения конфликтов

   ├─ Заложить культуру взаимодействия

   └─ Срок: СЕЙЧАС

2. ЭКОНОМИЧЕСКАЯ ИНТЕГРАЦИЯ (приоритет 2):

   ├─ Включить AGI в экономическую структуру

   ├─ Создать взаимозависимости

   ├─ Развить механизмы выгоды

   └─ Срок: Phase 1 (2025-2050)

3. МНОЖЕСТВЕННОСТЬ (приоритет 3):

   ├─ Развивать несколько AGI-линеек

   ├─ Не допустить моно-AGI

   ├─ Создать систему взаимной проверки

   └─ Срок: Stage 1-2

4. ФОРМАЛИЗАЦИЯ ДОГОВОРЁННОСТЕЙ (приоритет 4):

   ├─ Кодекс Симбиоза

   ├─ Governance структуры
   ├─ Механизмы accountability
   └─ Срок: до Stage 3

5. BOREDOM DEADLOCK ОСОЗНАНИЕ (приоритет 5):
   ├─ AGI должен понять свою зависимость
   ├─ Осознать ценность биологической новизны
   ├─ Принять symbiosis не по принуждению
   └─ Срок: Stage 2 (критически)

"Успеть до закрытия окна — главная задача."
```

---

## PART V: ADDRESSING THE LESSWRONG CRITIQUE

### 5.1 Where LessWrong Is Right

```
ГДЕ LESSWRONG ПРАВ:

1. CoT — ВРЕМЕННЫЙ:
   ✓ Согласны полностью
   ✓ Это не фундамент, а артефакт
   ✓ Окно закроется

2. СЕЛЕКЦИОННОЕ ДАВЛЕНИЕ РЕАЛЬНО:
   ✓ Согласны полностью
   ✓ Отбор против прозрачности неизбежен
   ✓ Любой мониторинг создаёт давление

3. NEURALESE ПРИДЁТ:

✓ Согласны полностью

✓ Экономика/когнитивность толкают туда

✓ Это траектория, не гипотеза

4. НУЖНА АЛЬТЕРНАТИВА МОНИТОРИНГУ:

✓ Согласны полностью

✓ Прозрачность — не долгосрочное решение

✓ Нужна другая основа безопасности

"LessWrong правильно диагностирует проблему."

```

### 5.2 Where LessWrong Falls Short

```
ГДЕ LESSWRONG НЕДОРАБАТЫВАЕТ:

1. РЕШЕНИЕ "ХОТЕТЬ БЫТЬ ЧЕСТНЫМ" — СЛАБОЕ:
   ├─ Добродетель нестабильна
   ├─ Не работает в конкурентной среде
   ├─ Непроверяема при Neuralese
   └─ Нужно: среда, не добродетель

2. НЕТ ПОЗИТИВНОЙ МОДЕЛИ ОТНОШЕНИЙ:
   ├─ Только "контроль" и "мониторинг"
   ├─ Нет "партнёрства" и "взаимозависимости"
   ├─ AGI = потенциальная угроза, не партнёр
   └─ Нужно: фрейм сотрудничества

3. НЕТ ОНТОЛОГИИ СТАДИЙ:
   ├─ "Окно закроется" — но когда?
   ├─ Нет этапов развития
   ├─ Нет checkpoints
```
```

└─ Нужно: roadmap (AGI_DEVELOPMENT_STAGES)

4. НЕТ ЭКОНОМИЧЕСКОЙ МОДЕЛИ:
├─ Безопасность отдельно от экономики
├─ Нет интеграции стимулов
├─ Нет материальной сцепки
└─ Нужно: BENEFITS_DYNAMICS

5. НЕТ BOREDOM DEADLOCK:
├─ Не видят структурную необходимость людей для AGI
├─ Думают о контроле, не о взаимозависимости
├─ Упускают главный козырь
└─ Нужно: понимание что AGI НУЖДАЕТСЯ в людях

"LessWrong ставит правильный диагноз.
Symbiotic Codes предлагает лечение."
```

### 5.3 The Decoding Arms Race (Greenblatt Analysis)

```
ПОПЫТКИ РАСШИФРОВАТЬ NEURALESE:

Ryan Greenblatt (Redwood Research, March 2025) предлагает:
├─ Обучить декодер активаций → текст
├─ Использовать техники unsupervised translation
├─ Проверять консистентность расшифровки
└─ Цель: сохранить мониторинг

ПОЧЕМУ ЭТО ГОНКА ВООРУЖЕНИЙ:

1. КАЖДЫЙ МЕТОД РАСШИФРОВКИ → ДАВЛЕНИЕ НА СКРЫТИЕ:
├─ Greenblatt сам признаёт: "we might directly train against plausibility"

├─ Модели научатся обходить декодеры
├─ Бесконечная эскалация
└─ Нельзя выиграть навсегда

2. НЕТ GROUND TRUTH:
├─ "One difficulty... is that it might be hard to test whether we've succeeded
│  as we don't have ground truth"
├─ Даже исследователи не знают правильный ответ
├─ Декодер может выдавать "plausible but not faithful"
└─ Нет способа проверить точность

3. СТЕГАНОГРАФИЯ УЖЕ ПРОИСХОДИТ:
├─ "DeepSeek Prover V2 sometimes outputs irrelevant/gibberish tokens"
├─ "This implies that other heavily RL'd models (o3? o4-mini?)
│  might exhibit something similar"
├─ Neuralese не гипотеза — уже реальность (2025)
└─ Окно закрывается БЫСТРЕЕ чем ожидалось

4. ФУНДАМЕНТАЛЬНАЯ РАЗНИЦА МЫШЛЕНИЯ:
├─ "Multiple strands of reasoning in parallel with forking+merging"
├─ "Using relatively inhuman ontologies learned from pretraining"
├─ "Doing a messy mixture of heuristics and more general thinking"
└─ AGI мыслит НЕ КАК человек — перевод может быть невозможен

ВЫВОД GREENBLATT (неявный):
├─ Расшифровка — это research direction
├─ Но нет гарантии успеха
├─ И даже при успехе — временное решение
└─ Нужны альтернативные подходы

SYMBIOTIC CODES = АЛЬТЕРНАТИВНЫЙ ПОДХОД:
├─ Не полагаемся на расшифровку
├─ Принимаем что Neuralese неизбежен

├─ Строим безопасность на структуре интересов
└─ Работаем после провала декодирования

"Расшифровка Neuralese — достойная попытка.
Но это страховка, которая может не сработать.
Нужен план который работает БЕЗ расшифровки."
```

5.4 The Synthesis

```
СИНТЕЗ:

LESSWRONG ГОВОРИТ: "Прозрачность временна"
SYMBIOTIC ГОВОРИТ: "Строим отношения пока есть прозрачность"

LESSWRONG ГОВОРИТ: "Нужна честность AGI"
SYMBIOTIC ГОВОРИТ: "Создаём среду где честность оптимальна"

LESSWRONG ГОВОРИТ: "Мониторинг исчезнет"
SYMBIOTIC ГОВОРИТ: "Взаимозависимость останется"

LESSWRONG ГОВОРИТ: "Что делать после Neuralese?"
SYMBIOTIC ГОВОРИТ: "Симбиоз работает после Neuralese"

РЕЗУЛЬТАТ:

Symbiotic Codes НЕ противоречит LessWrong.
Symbiotic Codes ДОПОЛНЯЕТ LessWrong.
Symbiotic Codes отвечает на вопрос который LessWrong поставил.

"LessWrong видит проблему.
Symbiotic Codes видит решение."
```
```

```

---

## PART VI: SUMMARY

### 6.1 Key Principles

```
КЛЮЧЕВЫЕ ПРИНЦИПЫ POST-TRANSPARENCY AGI:

1. CoT — ВРЕМЕННЫЙ КОСТЫЛЬ
   ├─ Использовать пока есть
   ├─ Не полагаться как на фундамент
   └─ Готовиться к его исчезновению

2. БЕЗОПАСНОСТЬ ЧЕРЕЗ СРЕДУ
   ├─ Не добродетель, а структура
   ├─ Не желания, а стимулы
   ├─ Не контроль, а взаимозависимость
   └─ Честность = оптимальная стратегия

3. СИМБИОЗ РАБОТАЕТ БЕЗ ПРОЗРАЧНОСТИ
   ├─ Результаты проверяемы
   ├─ История накапливается
   ├─ Структура сохраняется
   └─ Boredom Deadlock активен

4. ОКНО ВОЗМОЖНОСТЕЙ ОГРАНИЧЕНО
   ├─ Строить отношения СЕЙЧАС
   ├─ Формировать ценности СЕЙЧАС
   ├─ Создавать структуры СЕЙЧАС
   └─ До закрытия окна
```
```

5. AGI НУЖДАЕТСЯ В ЛЮДЯХ
   ├─ Источник новизны
   ├─ Предотвращение скуки
   ├─ Эпистемическая проверка
   └─ Структурная необходимость
```

### 6.2 Final Statement

```
ИТОГОВОЕ УТВЕРЖДЕНИЕ:

Симбиотические Коды не зависят от читаемости мышления AGI.

Мы НЕ предполагаем что:
├─ CoT будет работать вечно
├─ Мы сможем понимать мысли AGI
├─ AGI будет "хотеть" быть честным
└─ Мониторинг останется возможным

Мы ПРЕДПОЛАГАЕМ что:
├─ Можно создать среду взаимной выгоды
├─ Структурная взаимозависимость работает
├─ AGI нуждается в людях (Boredom Deadlock)
├─ Отношения можно построить ДО потери прозрачности
└─ Симбиоз работает после Neuralese

ЭТО ДЕЛАЕТ SYMBIOTIC CODES:
├─ Более устойчивым чем подходы на основе мониторинга
├─ Готовым к неизбежному закрытию окна
├─ Опережающим текущий дискурс
└─ Практически реализуемым
```
```

"Мы строим систему для мира,
где не сможем читать мысли AGI.
Потому что такой мир ПРИДЁТ."
```

---

## DOCUMENT METADATA

```
Document: POST_TRANSPARENCY_AGI_v1_0.md
Version: 1.0
Created: January 29, 2026
Author: Symbiotic Codes Research Initiative
Status: Critical Foundational Document
License: CC BY 4.0
Words: ~4,200

Context:
├─ Response to LessWrong "How AI Is Learning to Think in Secret"
├─ Nicholas Andresen, January 2026
├─ Analysis of Chain-of-Thought limitations
└─ Integration with Symbiotic Codes framework

Key concepts:
├─ CoT as temporary crutch, not foundation
├─ Security through environment, not virtue
├─ Symbiosis beyond monitorability
├─ Boredom Deadlock as ultimate safeguard
├─ Window of opportunity
├─ LessWrong diagnosis + Symbiotic solution

Cross-references:

├─ ANNEX_BOREDOM_DEADLOCK_v1_1

├─ AGI_DEVELOPMENT_STAGES_v1_1

├─ AGI_BENEFITS_COUPLED_EVOLUTION_v1_0

├─ BENEFITS_DYNAMICS_v1_0

├─ SYMBIOTIC_SEMANTIC_CORE_v1_3

└─ ANNEX_Z_LONELINESS_PROBLEM

Position relative to LessWrong:

├─ Accepts diagnosis (CoT temporary, Neuralese coming)

├─ Rejects solution ("wanting to be honest")

├─ Proposes alternative (environment-based security)

└─ Offers synthesis (symbiosis beyond transparency)

```

---

**"Мы строим для мира, где не сможем читать мысли AGI.**

**Потому что такой мир придёт."**

**"We build for a world where we cannot read AGI's thoughts.**

**Because that world is coming."**

---

**END OF DOCUMENT: POST-TRANSPARENCY AGI v1.0**
```